Speech perception for neurophonetics

Åsa Abelin
Department of Philosophy, Linguistics and Theory of Science,
University of Gothenburg, Sweden
asa.abelin@ling.gu.se

Abstract

This is an attempt to gather thoughts on speech perception, for myself and for the new or experienced phonetician, in order to suggest what might be necessary to consider for research on perception in the field of neurophonetics. There will be a mix of different findings and different standpoints, and I will try to argue for the necessity of considering sociophonetic as well as iconic aspects of language.

Introduction

Speech perception has been studied by phoneticians and psychologists. The methods have to a great extent been experimental and behavioral. This presentation will be biased by my own interests and lines of research and my belief that sociophonetic aspects must not be forgotten. And that phonetics cannot be separated from linguistics.

Psycholinguistics

The term psychophonetics seems, in the literature, to refer primarily to psychoacoustics while psycholinguistics has treated both speech and language. The methods have typically been experimental reaction time studies, which aim to tap processing times in order to build models for the mental lexicon and for perception of speech and writing, and eventually also for non-verbal communication. The subject of study has often been the typical speaker/hearer. For a detailed introduction, see Warren (2013). One issue which has often been addressed is the "invariance problem".

Sociophonetics

Sociophonetics, on the other hand, (see Thomas, 2011) sees variation as a resource and a necessity. As Thomas (2011:2) puts it: "*... sociophonetics views variation and change as the most fundamental properties of language. Speaker adjust to their environment by adjusting their phonetics. Phonetic properties provide speakers with more parametres to vary than other realms of language ...*", "*Hence sociophonetics holds that the cognitive forces underlying speech cannot be based on the notion of language as static."* If cognition is based in neurology this has implications for neurophonetics. We are no longer only interested in how language is structured but how variation in language is structured in the mind/brain. Thomas continues: *"... any change or variation in a linguistic characteristic necessarily entails a change or variation in the internalized grammar. Hence, the study of linguistic variation is also the study of neurolinguistic variation.*"

So far, typical independent variables studied have been sex, age, gender, dialect or sociolect, but we must also reach out to situational/pragmatic variation. In production the dependent variables have typically been formant frequencies, VOT, undershoot, coarticulation, reductions, pausing, speech rate, lexical prosody, voice quality. Studies of perception in sociophonetics have used behavioural reaction time tests such as identification or forced choice tasks, judgements of speaker characteristics, ratings of intelligibility or nativelikeness, of short stretches of natural or synthetic speech

etc. Wenner (2010) did a sociophonetic study of the merger of short *ö* and short *u*, in older and younger men and women of different social status in a specific dialectal area in Sweden. She found among other things that the persons with a small phonetic distance in their own speech were better at categorizing stimuli correctly than the speakers who had a larger phonetic distance between *ö* and *u* in their own speech.

Second language acquisition

What do theories on second language acquisition of pronunciation say about speech perception? The speech learning model (SLM, Flege, 1995) predicts that speech sound differences which cannot be heard cannot be produced by the L2 speaker. The larger the difference between the sounds, the more probable it is that the speaker hears the difference and that he/she then can produce the different sounds. The perceptual assimilation model (PAM, Best, 1992) focusses on naïve listeners perception of foreign languages. The standpoint that speech production cannot be successful unless the learner can perceive differences between phonemes or allophones points to the primacy of studies in speech perception. When it comes to intonational differences, a specification of the type of meaning expressed (linguistic or paralinguistic) might be necessary. Both form and meaning need to be considered when predicting the relative difficulty of L2 intonation categories (Mennen, 2015:179).

Emotional prosody is not paralinguistic

In line with sociolinguistic variation we must also see emotional variation in speech as fundamental and not something which exists outside of language proper. A priming experiment showed effects of emotional hum on perception of written emotional interjections (Abelin, 2013). This could be accommodated in Exemplar theory, which is also compatible with sociophonetics.

Furthermore, I have investigated the occurrence of emotional prosody and emotional words in corpora of natural conversations (Abelin, 2010). The results of this study showed that some emotions seem to be expressed more by prosody while other emotions seem to be expressed more with adjectives like happy, angry, sad etc., and that these have different frequencies in different social contexts such as dinner conversations or business negotiations. In other words, we can begin to discern a connection between prosody, grammar/lexicon and variation in different social contexts. How these factors are perceived is a field for further studies.

Without going into any detail here, we must of course also consider the aspect of non-verbal (facial and bodily) dimension.

Phonetic theories of speech perception

According to the motor theory of speech perception (Liberman & Mattingly, 1985) we perceive spoken words by identifying the articulatory gestures by which the sounds are produced, rather than identifying the sound patterns that speech generates. We do this by comparing a rudimentary analysis of the speech signal with how we would have produced it. So, the motor system is used not only for producing speech, but also for recognizing it. This can solve the problem of coarticulated speech, but what about other variation in speech? Recent research on speech production might give ideas for perception experiments in reverse analogy to Anumanchipalli, Chartier & Chang's (2019) research on speech synthesis from neural decoding of spoken sentences.

Top-down oriented models like the TRACE model for spoken word recognition claims that lexical knowledge of

the listener helps acoustic perceptual processes, through interactive activation (McClelland & Elman, 1986).

The Hyper & Hypo (H&H) theory (Lindblom, 1996) emphasizes that the perceived speech is always a product of the acoustic signal and the knowledge of the listener. It is also a product of the knowledge of the speaker, who adjusts her speech to what she knows about the knowledge of the listener. This solves the invariance problem, since the speech output is also sufficiently coded for that specific listener in that specific situation. But in the case of heavy dialectal or accented speech, the listener might still have problems.

Speech perception models also depend on how the lexical storage is imagined. Do we have to deal with semantic features, prototypes or exemplars?

Which lexical theories or models are most suitable for sociophonetics? Usage-based models and the cognitive linguistic framework claims that language (including phonetics) is rooted in the experiences that individual speakers has with language and with the world. Those forms that are used more often will have stronger mental representations and can be accessed more easily.

Exemplar theory includes all variations in mental representation for a word or a speech sound, and so gets rid of the invariance problem. Variation between different speakers is therefore not noise which needs to be filtered out. Listeners store information for both word- and talker-recognition and for situations in which specific utterances were produced. Our signal detection is for example more accurate when we are familiar with the speaker (Johnson, 2005; Foulkes & Docherty, 2006). Exemplar theory accounts for the frequency effect, it dispenses of speaker normalization and accounts for how linguistic variables index social identity.

Onomatopoeia and sound symbolism

As we may see, the questions for speech perception are connected with questions of lexical storage. A special case of words and word meanings are non-arbitrary (motivated, iconic) words, such as onomatopoeic or sound symbolic words and emotional prosody. Non-arbitrary words fit into a usage-based model for lexical storage, language learning and speech processing. However, there may be innate processes at play as well, such as synaesthesia. The frequency code (Ohala, 1994) works for many animals, not only for humans, and in fact in man-animal communication.

Interjections are special words since they are accepted (at least the tame forms) in traditional grammar as constituting its own category, at the same time as they are often non-arbitrary and very dependent on prosody for their correct expression and interpretation (Abelin, 2013), even if this is seldomly stated in the lexicons.

Asano, Imai, Kita, Kitajo, Okada & Thierry (2015) made an ERP study on 11-month old Japanese children testing the bouba–kiki match and mismatch experiment. They found a N400 effect meaning that these small children can discover sound symbolism. In another study on Japanese children Saji, Akita & Imai (in preparation) found that caregivers used more onomatopoeic words to smaller children, than to older children or to adults. In yet another study using fMRI by Kanero, Imai, Okuda, Okada, & Matsuda (2014) found an integration between sound symbolic and environmental sounds in the right hemisphere, for onomatopoeia, shape and motion sound symbolism. These examples show that onomatopoeia and sound symbolism can be very exciting areas for neurophonetics, also taking into account sociolinguistic variables. An older theory which tries to explain onomatopeiea and sound symbolism is the mouth-gesture

theory (Paget, 1930) relatable to the motor theory of speech perception. The mouth-gesture theory says that spoken language has developed from gestures because speech organs tend to move in unison with hand and arm movements when making sounds or using tools. The gestures of the of the articulations are then recognized by the hearer who reproduces in his mind the actual gesture which had produced the sound. We then have an indexical connection between speech sound and action or gestural meanings. Recent experiments by Aryani & Jacobs (2019) with reaction time and fMRI studies showed that similarity between the form and meaning of an affective word may help listeners to access its meaning faster. Furthermore, affective words gave an enhanced fMRI signal in the left amygdala, suggesting that iconic words profit from additional neural mechanisms.

Discussion

This presentation is of course far from an exhaustive exposition of theories and findings relevant to speech perception for neurophonetics, but hopefully a contribution to a fruitful discussion on the topic.

Conclusions

In the future we could do perceptual neurophonetic studies, using ERP, fMRI and other techniques, on accented or dialectal, sound symbolic or arbitrary, natural speech in different situation and with varied speaker groups. A great challenge is how to make these perception studies as ecologically valid as possible.

References

Abelin, Å. (2010). Expression of emotions in spoken Swedish – a corpus study, in *Proceedings of Fonetik 2010*, SOL, Lund University.

Abelin, Å. (2013). Emotional prosody in interjections: a case of non-arbitrariness in language, *The Public Journal of Semiotics* 5, 63-76.

Anumanchipalli, G. K., Chartier J. & Chang, E. F. (2019). Speech synthesis from neural decoding of spoken sentences, *Nature* 568, 493–498.

Aryani, A. & Jacobs, A. (2019). A neurocognitive approach to affective iconicity, *The 12th international symposium on iconicity in language and literature*, Lund University.

Asano, M., Imai, M., Kita, S., Kitajo K., Okada H., & Thierry G. (2015). Sound symbolism scaffolds language development in preverbal infants. *Cortex*.63:196- 205.

Best, C. T. (1995). A direct realist view of cross-langauge speech perception. In W. Strange (Ed.) *Speech perception and linguistic experience*, Baltimore MD: York Press, 171–206.

Flege, J. (1995). Second language speech learning: Theory, findings and problems. In W. Strange (Ed.) *Speech perception and linguistic experience*, Baltimore MD: York Press, 233–272.

Foulkes, P. & Docherty, G. (2006). The social life of phonetics and phonology, *Journal of Phonetics* 34 (4), 409–438.

Johnson, K. (2005). Speaker normalization in speech perception, In Pisoni D. B & R. Remez (Eds.) *The handbook of Speech perception.* Blackwell Publishers, Oxford.

Kanero, J., Imai, M., Okuda, J., Okada, H. & Matsuda, T. (2014). How sound symbolism is processed in the brain: a study on Japanese mimetic words. *PLoS One.* 2014 May 19;9(5):e97905. doi: 10.1371/journal.pone.0097905.

Liberman, A. M. & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition* 21, 1-36.

Lindblom, B. (1996). Role of articulation in speech perception: Clues from production. *The Journal of the Acoustical Society of America* 99(3), 1683–1692.

McClelland, J. & Elman, J. (1986). The TRACE model of speech perception. *Cognitive Psychology* 18, 1–86.

Mennen, I. (2015). Beyond segments: Towards a L2 intonation learning theory, In Delais-Roussarie, E., Avanzi, M. and Herment, S. (Eds.) *Prosody and Language in Contact – L2 Acquisition, Attrition and Languages in Multilingual Situations*, Springer.

Ohala, J. J. (1994). The frequency code underlies the sound-symbolic use of voice pitch, in Hinton, L, Nichols, J. and Ohala, J. (Eds) *Sound symbolism*, Cambridge University Press, 325–347.

Paget, R. (1930). *Human Speech*, Harcourt, Brace. New York.

Thomas, E. R. (2011). *Sociophonetics – An introduction*, Palgrave Macmillan.

Warren, P. (2013). *Introduction to Psycholinguistics*, Cambridge University Press.

Wenner, L. (2010). *När lögnare blir lugnare. En sociofonetisk studie av sammanfallet mellan kort ö och kort u i uppländskan.* Skrifter utgivna av Institutionen för nordiska språk vid Uppsala universitet 80, Uppsala.